micromedia
a division of IHS Canada

20 Victoria Street
Toronto, Ontario M5C 2N8
Tel: (416) 362-5211
Toll free: 1-800-387-2689
Fax: (416) 362-6161
Email: info@micromedia.on.ca

OQRE

Office de
la qualité et
de la responsabilité
en éducation

OQRE Série d'études de recherche

• NUMÉRO 2 • JUIN 1999

MISSION DE L'OQRE

L'OQRE assurera une plus grande responsabilité et contribuera à améliorer la qualité de l'éducation en Ontario. Cela s'effectuera au moyen d'évaluations basées sur des données objectives, fidèles et pertinentes et de la publication en temps voulu de ces données et de la formulation de recommandations aux fins d'amélioration du système.

La présente étude a été écrite sous contrat pour l'Office de la qualité et de la responsabilité en éducation (OQRE). Les opinions exprimées sont celles des auteur(e)s et ne sont pas nécessairement celles de l'OQRE.

Office de qualité et de la responsabilité en éducation
2, rue Carlton, bureau 1200
Toronto (Ontario) M5B 2M9

Téléphone : 1-888-327-7377

Site Web : www.eqao.com

ISBN 0-7778-9133-6

Avant-propos

L'OQRE croit que les évaluations à grande échelle peuvent contribuer à apporter des changements constructifs au système d'éducation si elles amènent les éducatrices et les éducateurs, les parents et les élèves à réfléchir à ce qui se passe dans la classe et à en discuter.

Le mandat de l'OQRE est fondé sur deux engagements clés : la responsabilité et l'amélioration. Nous faisons rapport aux parents, aux éducatrices et aux éducateurs et au public sur le rendement des élèves et la qualité de l'éducation dans le système scolaire financé par les deniers publics. Nous veillons également à ce que ces renseignements servent de catalyseur pour améliorer l'enseignement et l'apprentissage.

L'une des façons de favoriser l'amélioration est de soutenir la recherche sur les facteurs qui ont une influence sur le rendement des élèves et la qualité de l'éducation. C'est ce que fait l'OQRE avec sa *Série d'études de recherche* qui fournit des analyses et des renseignements sur un vaste éventail de questions en rapport avec l'éducation.

Cette série a deux objectifs importants. Le premier est d'augmenter nos connaissances sur ce qui donne de bons résultats dans la classe, dans l'école et dans le conseil scolaire. Le second est de combler l'écart entre ce que nous savons et ce que nous faisons. Les recherches peuvent avoir un impact positif et durable sur l'apprentissage des élèves si elles donnent aux éducatrices et aux éducateurs les connaissances, les occasions et la motivation nécessaires pour adapter et affiner leurs stratégies et leurs façons de procéder sur la base de données exactes et fiables. L'OQRE soutient toutes les formes d'études systématiques qui favorisent le changement constructif et l'amélioration.

J'ai le plaisir de présenter cette dernière étude de la *Série d'études de recherche de l'OQRE.*

La directrice générale,

Joan M. Green
Juin 1999

TABLE DES MATIÈRES

Différents types d'évaluation : Validité et fiabilité

Anthony W. Bartley
Université Lakehead

Alexandra Lawson
Institut d'études pédagogiques de l'Ontario
Université de Toronto

On a assisté, ces dernières années, à un déplacement fondamental dans la nature des évaluations à grande échelle et dans l'utilisation qui en est faite. Le changement a été marqué par le remplacement des évaluations fondées uniquement sur des questions à réponse choisie par des évaluations fondées sur la mesure directe des résultats attendus des élèves. Le présent article discute des problèmes posés par l'utilisation de certaines de ces nouvelles formes d'évaluation, particulièrement d'un point de vue technique. L'accent est mis sur la validité et la fiabilité des évaluations complémentaires, avec une attention particulière sur les théories qui les sous-tendent. Différents types d'évaluation y sont analysés.

L'utilisation des évaluations à grande échelle a continué à se répandre en Amérique du Nord au cours des trente dernières années. Leur objectif déclaré a généralement été d'évaluer et d'améliorer la qualité de l'enseignement et de l'apprentissage (Stake, 1998). Les évaluations à grande échelle de la première génération étaient considérées comme des indicateurs neutres des progrès des élèves sans rapport direct avec ce qui se passait dans la classe (Cole, 1991, p. 97-98). Mais, lorsque l'on a commencé à mettre l'accent sur ce qu'on a appelé les « savoirs de base », au cours des années 1970, et que des programmes d'évaluation fondés sur des critères qui assignaient des notes aux conseils scolaires et aux écoles ont été établis, ces indicateurs « neutres » sont devenus partie intégrante du système d'éducation. Lovitts et Champagne (1990) avancent que ce type d'évaluation à grande échelle basée sur des questions à réponse choisie répondait davantage aux besoins des personnes responsables des prises de décisions qu'à ceux des titulaires de classes et de leurs élèves, en partie parce qu'il produit des données d'ensemble susceptibles de manipulation statistique. Cela a permis aux responsables politiques et aux cadres supérieurs de suivre la situation des programmes dans un district, un État, une province, ou même un pays, comme dans le cas du *National Assessment of Educational Progress* (NAEP). Des voix de plus en plus nombreuses commencèrent à s'élever contre ces pratiques que l'on accusait d'aller à l'encontre de l'apprentissage authentique au lieu de le favoriser. Il devint évident que les enseignantes et les enseignants sous pression se concentraient sur la préparation de l'évaluation et passaient davantage de temps sur les sujets couverts par les tests et moins de temps sur les « extra ». Elles et ils avaient tendance à préparer leurs élèves aux tests en insistant sur la mémorisation et le travail forcé sur des données fragmentées (Smith, 1991). De plus, elles et ils avaient souvent recours à des moyens douteux pour rehausser les résultats en « aidant » les élèves le moment venu ou en les exerçant à des techniques de préparation des tests, toutes choses qui « polluent » les résultats (Nolen, Haladyna et Hass, 1992; Cannell, 1988).

En réponse à ces problèmes, un grand nombre de réformatrices et de réformateurs ont suggéré d'utiliser différents types d'évaluations complémentaires (Shepard, 1989; Madaus et Kellaghan, 1993; Worthen, 1993). Elles et ils soutenaient qu'il fallait modifier aussi bien les tests eux-mêmes que la façon dont ils étaient administrés de manière à favoriser à la fois l'enseignement et l'apprentissage. Wiggins (1989) déclare qu'il faut concevoir des évaluations qui établissent des normes et enseigner en conséquence, de sorte que la préparation et l'administration des tests favorisent effectivement l'éducation au lieu de la compromettre (p. 41). Un certain nombre d'autres évaluations complémentaires ont été suggérées, comme l'évaluation authentique, l'évaluation à partir du dossier d'apprentissage ou les tâches fondées sur la performance de l'élève. Ces méthodes partagent toutes un certain nombre de caractéristiques fondamentales : les tests sont notés par des gens et non par des machines, ils requièrent une réflexion plus élevée,

ils sont similaires aux bonnes pratiques d'enseignement en classe, ils se réfèrent à des critères plutôt qu'à des normes, et les tâches qui les constituent sont des activités d'apprentissage intelligentes (Herman, Aschbacher et Winters, 1992). Par ailleurs, ils donnent aux enseignantes et aux enseignants des renseignements sur leurs méthodes d'enseignement et leur efficacité.

Un grand nombre de systèmes scolaires ont répondu en adoptant une variété d'autres types de tests à utiliser non seulement par les enseignantes et les enseignants mais aussi dans les programmes d'évaluation à grande échelle. Cependant, la plupart des systèmes qui ont tenté de réformer leurs programmes d'évaluation en y incorporant de nouvelles techniques se sont heurtés à toutes sortes de graves difficultés tout au long du processus de conception et d'application (Kirst et Mazzeo, 1996; Jones et Whitford, 1997). Notre présent propos est d'examiner seulement deux des problèmes techniques qui se sont posés : la validité et la fiabilité des évaluations complémentaires utilisées dans le cadre des programmes d'évaluation à grande échelle.

L'évaluation en Ontario

Au cours des années 1990, l'Ontario a affiché un intérêt accru pour les évaluations à grande échelle, a participé aux évaluations existantes et en a élaboré de nouvelles. D'importants échantillons d'élèves en provenance de toute la province ont participé à des projets nationaux comme le Programme d'indicateurs du rendement scolaire (PIRS) en lecture, écriture, sciences et mathématiques, et la Troisième enquête internationale sur les mathématiques et les sciences (TEIMS). Des tests ont en outre été administrés à toutes et tous les élèves de 9e année en lecture et en écriture en 1993-1994 et 1994-1995. Ces évaluations ont été effectuées sous l'égide du ministère de l'Éducation et de la Formation, comme tous les tests appliqués précédemment en Ontario.

La participation de l'Ontario à ces projets à grande échelle était bien connue dans les rangs du personnel d'administration de l'éducation publique, mais c'est le rapport de la Commission royale sur l'éducation (1994) qui a généralisé les observations sur la nécessité d'adopter une politique sur l'évaluation en Ontario :

> Le fait que beaucoup de personnes demandent, avec une certaine impatience et insistance, des renseignements plus complets sur les résultats des élèves dans tout l'Ontario témoigne de la pénurie de ce genre de renseignements depuis plusieurs décennies, comparativement à des temps anciens et à d'autres endroits du monde, et reflète l'actuelle « crise de confiance » touchant l'éducation, inquiétude certainement alimentée par le manque de données concrètes. (p. 161)

Le rapport demandait que le gouvernement de l'Ontario et les associations et organismes professionnels du secteur de l'éducation « prennent un engagement ferme et durable en matière d'évaluation pour améliorer la situation et assumer pleinement leurs responsabilités envers la population » (Commission royale, 1994, p. 165). Bien que nous n'ayons pas l'intention de répéter ici le rapport de la Commission royale, nous tenons à rappeler certaines de ses importantes déclarations et recommandations. Les commissaires ont décrit de la façon suivante le test de 9e année en lecture et en écriture :

> C'était en réalité un très bon test : premièrement, il durait plus de six heures, réparties sur deux semaines, ce qui permettait aux élèves de montrer ce qu'ils avaient acquis et compris d'une façon qui aurait été impossible avec un test type d'une heure sur les « savoirs de base » ou un examen comparable. Deuxièmement, le test consistait non seulement en des questions à réponse brève, mais en une véritable évaluation du rendement. [...] Nous applaudissons aux efforts déployés par le Ministère pour organiser des tests à grande échelle visant à évaluer l'apprentissage de manière authentique. (p. 181)

C'était là une déclaration sans équivoque en faveur d'un type d'évaluation différent. Si l'on y ajoute la recommandation n° 50 sur l'administration d'un test en lecture, écriture et mathématiques aux élèves de 3e année et la recommandation n° 51 sur l'établissement de l'Office de la redevabilité et du rendement de l'apprentissage, maintenant l'Office de la qualité et de la responsabilité en éducation (OQRE), le programme d'évaluation de l'Ontario était lancé.

Les tests administrés par l'OQRE ont été conçus avec deux objectifs : fournir des renseignements en réponse aux exigences de responsabilité et favoriser l'amélioration de l'apprentissage chez toutes et tous les élèves (OQRE, 1997). Fort de ce mandat, l'OQRE a commencé à élaborer des instruments d'évaluation susceptibles non seulement de fournir des données techniquement fiables sur le rendement des élèves par rapport au curriculum en vigueur, mais également de renforcer l'enseignement et l'apprentissage. Pour que soient générés des renseignements crédibles et défendables, les personnes chargées de concevoir les instruments d'évaluation devaient également aborder les questions de validité et de fiabilité.

Considérations techniques :

Validité

À l'origine, la discussion des questions techniques posées par la conception et l'administration des évaluations à grande échelle était limitée au personnel de recherche et au milieu de l'évaluation en éducation. Les *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association et National Council on Measurement in Education, 1985) continuent à présenter les questions techniques d'une façon quelque peu mécaniste. Les *Standards* ont été rédigés avant que le mouvement en faveur de l'évaluation de la performance n'ait pris de l'ampleur. Au cours des deux dernières décennies, cependant, la notion de validité a subi des modifications notables qui témoignent d'un regroupement d'éléments préalablement isolés comme le contenu, les critères ou la validité (AERA, APA et NCME, 1985). Comme le soutient Messick (1989), la validité ne peut plus être séparée en secteurs distincts mais doit être vue comme

> un jugement d'évaluation intégré de la mesure dans laquelle les preuves empiriques et les principes théoriques soutiennent la justesse et la pertinence des inférences et des actions basées sur les résultats des tests et d'autres modes d'évaluation. (p. 18)
>
> [C'est nous qui traduisons.]

Avec cette définition, le travail de validation devient un « jugement d'évaluation intégré » où l'on fait appel à la fois aux preuves et aux principes pour examiner la « valeur fonctionnelle » (Messick, 1989, p. 18) de l'évaluation. Les théories à examiner auraient tendance à porter sur ce qui suit :

Raison d'être : Quels sont les objectifs visés par l'évaluation?

Nature des questions utilisées pour l'évaluation : Quels sont les types de questions les plus appropriés pour cette évaluation?

Apprentissage : Le test a-t-il été élaboré en fonction des théories de l'apprentissage appropriées pour chaque secteur du programme d'études?

Équité et biais : Les questions dans leur ensemble ont-elles pour effet de discriminer injustement en faveur d'un groupe particulier ou contre lui?

La façon dont les questions représentent le curriculum : Faut-il donner la même importance à tous les domaines?

Le système de notation : Comment chaque question devrait-elle être notée et comment les notes assignées aux questions individuelles doivent-elles être combinées pour donner la note finale (et publiée)?

Résultats publiés et potentiel de généralisation : Que révèlent les résultats des élèves au test sur leur rendement possible à l'égard de l'ensemble du programme d'études?

Conséquences : Quel a été l'effet de l'évaluation sur les personnes et sur les organismes? Cet effet soutient-il les objectifs visés?

Ces théories peuvent être associées à trois phases distinctes de l'évaluation :

(1) la conception du test;

(2) l'administration et la notation du test;

(3) le rapport sur le test et ses conséquences.

La validation doit également tenir compte des preuves empiriques. Ces preuves seraient recueillies à l'issue des mises à l'essai puis de l'administration du test lui-même. Les secteurs examinés auraient tendance à comprendre ce qui suit :

Réponses aux questions : Les élèves ont-elles et ils été en mesure de donner des réponses qui montraient qu'elles et ils comprenaient ce qui était demandé?

Durée de l'évaluation : Les élèves ont-elles et ils pu terminer les tâches pendant le temps qui leur était alloué? Y avait-il un ralentissement de la performance vers la fin de la période consacrée au test?

Potentiel de généralisation et cohérence interne : Les élèves ont-elles et ils travaillé de façon uniforme pendant tout le test? Faut-il s'attendre à ce qu'elles et ils le fassent?

Considérations techniques :

Validité

Utilisation du système de notation : Le système de notation était-il adéquat? Les enseignantes et les enseignants chargés de la notation ont-elles et ils bien compris le système? Les tâches ont-elles pu être notées dans les délais prévus?

Crédibilité : Les résultats publiés sont-ils crédibles compte tenu des preuves du rendement des élèves déjà recueillies par les parents et les enseignantes et enseignants?

Le problème que pose la validation, c'est qu'une fois que chacune de ces sources de preuves a été examinée et que les théories sous-jacentes ont été étudiées, il faut prendre des décisions sur ce qui peut être revendiqué et publié sur le rendement des élèves (au niveau de l'individu, de la classe ou du conseil scolaire). En outre, on est de plus en plus préoccupé par ce que l'on reconnaît comme les conséquences sociales des données fournies par les évaluations (Gipps, 1994). La question n'est pas seulement de décider si les résultats constituent une interprétation valide de la situation, mais aussi de juger si l'évaluation devrait être utilisée aux fins auxquelles elle est prévue. Bien que le débat se poursuive sur ce qui constitue le poids relatif approprié de la validité corrélative (Brown, McCallum, Taggart et Gipps, 1997; Mehrens, 1992), cette importante question reste posée. Pour poursuivre l'examen de la validité et le mettre en contexte, nous devons également examiner sa contrepartie, la fiabilité.

Considérations techniques :

Fiabilité

Dans le domaine de l'évaluation, la fiabilité renvoie à la cohérence ou à la non-cohérence des mesures (Feldt et Brennan, 1989). La théorie classique de la fiabilité repose sur les notions d'erreur-type, de cœfficient de fidélité et de note vraie; elle a beaucoup été utilisée dans l'analyse des tests normatifs à réponse choisie. Mais la théorie classique de la fiabilité s'avère limitée lorsqu'elle est appliquée à des évaluations de la performance où les élèves disposent d'une importante latitude dans la façon dont elles et ils interprètent les tâches et y répondent, et où les tâches sont moins nombreuses, mais plus longues.

À propos de la fiabilité, Traub et Rowley affirment que les mesures qui manquent de cohérence sont le fléau de la recherche (1991, p. 37). Bien que cela puisse représenter un problème fondamental dans l'analyse des évaluations fermées où il n'y a qu'une seule « bonne réponse », il y a beaucoup d'exemples de tâches d'évaluation en écriture, mathématiques, sciences ou beaux-arts où les élèves sont invités à planifier, concevoir et essayer toute une gamme de réponses possibles. Il est particulièrement intéressant de noter que le problème revient à une discussion de la validité du niveau d'incohérence attendu et des théories à utiliser pour expliquer à la fois les cohérences et les incohérences. La fiabilité est vue comme une condition nécessaire mais non suffisante de la validité de l'interprétation des données. Dans le manuel intitulé *Educational Measurement*, Feldt et Brennan (1989) mettent en garde contre un souci exagéré de la fiabilité. Ils reconnaissent

> l'importance première de la validité dans l'évaluation de l'exactitude d'une mesure éducationnelle. Aucun ensemble de données de fiabilité, quelles que soient les mesures utilisées pour l'analyser, n'a de valeur si la mesure à laquelle il s'applique est hors de propos ou redondante. (p. 143)
>
> [C'est nous qui traduisons.]

Quelle valeur faut-il alors accorder aux preuves de fiabilité? Si le test présente un nombre considérable de questions à réponse choisie, on peut envisager d'utiliser les méthodes classiques, test et retest et corrélations pair-impair, où la preuve consiste à retrouver les mêmes résultats d'un test à l'autre. Ces méthodes risquent de ne pas être appropriées, cependant, pour les questions à réponse élaborée. Pour le type d'évaluation utilisé par l'OQRE, il faut modifier le sens que l'on donne à la notion de fiabilité et à ses rapports avec la validité (Gipps, 1994). Ceci dit, nous avons quand même besoin d'avoir la preuve, tout d'abord, que les enseignantes et les enseignants administrent le test d'une façon cohérente. C'est-à-dire que les procédures d'administration d'un test doivent permettre à chaque élève d'avoir la même possibilité de faire de son mieux sans permettre à aucun groupe d'élèves de jouir d'un avantage injuste. Lors de la formation des enseignantes et des enseignants et du personnel d'administration, il faudra clarifier la question d'intégrité et insister sur la nécessité de prendre le temps de présenter les tâches aux élèves et de répondre à leurs questions.

Nous sommes également préoccupés par la fiabilité de la notation, c'est-à-dire par l'objectivité des personnes qui en sont chargées. Cette question a fait l'objet d'une attention considérable dans les publications (Raymond et Houston, 1990; Slater et Ryan, 1993). La complexité des tâches qui donnent lieu à des réponses qualitativement différentes peut saper la fiabilité de l'interprétation des notes, comme dans les évaluations à partir du dossier d'apprentissage utilisées dans le Vermont (Koretz et al., 1992). Shavelson, Gao et Baxter, cependant, rapportent des résultats plus positifs pour des évaluations de performance en sciences et indiquent que les résultats sont cohérents. Ils indiquent que le cœfficient d'objectivité ne pose pas de problème et qu'il est possible d'apprendre aux évaluatrices et aux évaluateurs à noter la performance de façon fiable

en temps réel ou à partir de substituts, comme des cahiers (1994, p. 1). (Une étude plus détaillée de la fiabilité sera présentée dans le prochain numéro de la *Série d'études de recherche de l'OQRE.*)

Pour poursuivre sur un plan pratique l'examen de certaines de ces idées, nous allons maintenant passer en revue différentes formes et tâches d'évaluation utilisées dans les évaluations à grande échelle en lecture, écriture et mathématiques. Notre discussion, dans cette partie de l'article, s'articule autour de certaines questions précises sur la validation.

Validité des évaluations en lecture et en écriture

Question concernant la validité : *Le modèle d'évaluation correspond-il à l'objectif visé? Il s'agit de se demander si l'évaluation, comme instrument de responsabilité, donne une image exacte de l'habileté de l'élève dans un domaine donné.*

La demande d'élaboration de tests de lecture et d'écriture à administrer aux élèves du cycle primaire vient de l'inquiétude de la société devant le niveau de littératie des élèves. Cet intérêt généralisé a été à l'origine de l'établissement de programmes d'évaluation dans de nombreux territoires de compétence aux États-Unis, au Royaume-Uni et en Australie. Ces évaluations, cependant, ont peu de choses en commun, ni sur le plan de la conception, ni sur celui de l'administration; la différence entre les influences culturelles locales, en termes aussi bien du programme d'études que de l'évaluation, a donné lieu à des approches différentes. Horner (1998) compare les modèles de programme d'études :

> En Angleterre et au Pays de Galles, depuis les révisions récentes, les documents relatifs au programme d'études établissent, sous une forme succincte, le contenu, les habiletés et les processus qui doivent être enseignés. C'est là-dessus qu'est basée la description des niveaux à l'origine de la grille d'évaluation qui permet de mesurer le rendement des élèves à différentes étapes de leur scolarité. En d'autres termes, bien qu'il y ait des attentes quant à ce qui constitue le niveau « moyen » pour un groupe d'âge donné, la mesure est définie séparément des résultats visés par le programme d'études. Des modèles similaires sont en vigueur en Écosse et en Irlande du Nord. Par contre, les modèles utilisés en Australie, en Nouvelle-Zélande et aux États-Unis définissent les résultats attendus puis ce qui doit être fait pour que les élèves les atteignent. Ce modèle a beaucoup plus de chances d'être vu comme un mécanisme pour définir les compétences minimales alors qu'il peut sembler que le modèle du Royaume-Uni ne réussit pas à établir un lien essentiel entre ce qui est enseigné et ce qui est appris, puisque la différence est rendue évidente à la fin plutôt que pendant le processus d'enseignement du programme d'études. (p. 93)
>
> [C'est nous qui traduisons.]

L'application de cette comparaison du niveau de détail aux programmes-cadres récents de l'Ontario indique que les résultats visés par le *Programme d'études commun* (1995) sont moins clairement définis alors que

> *Le curriculum de l'Ontario, de la 1re à la 8e année, Français, 1997*, détermine avec précision et en détail pour chaque année d'études ce que les élèves doivent savoir et pouvoir faire. Ainsi, les conseils scolaires n'auront plus à rédiger leurs propres contenus d'apprentissage; le curriculum sera uniforme dans toute la province, ce qui facilitera la tenue de tests à l'échelle provinciale. (Ministère de l'Éducation et de la Formation, 1997, p. 4).

Le contraste entre l'ancien et le nouveau curriculum renvoie à la comparaison entre les méthodes utilisées au Royaume-Uni et dans les autres pays. Il y a aussi un soutien explicite, dans le nouveau curriculum, en faveur des évaluations à l'échelle de la province. En outre, *Le curriculum de l'Ontario, de la 1re à la 8e année, Français, 1997*, est, comme il le dit lui-même, notablement plus rigoureux et exigeant que les curriculum précédents. Les auteurs du curriculum soutiennent que cette rigueur vient de l'inclusion d'un éventail plus large de connaissances et d'habiletés et de l'introduction de nombreuses aptitudes dans les premières années d'études. C'est là où les justifications théoriques des changements apportés dans les premières années s'avéreraient particulièrement valables. Tyler (1949) a discuté il y a près de 50 ans de la pertinence d'une psychologie d'apprentissage en rapport avec le placement, dans les différentes années d'études, d'objectifs d'éducation réalisables (p. 38). À cette époque comme maintenant, on ne trouvait guère d'exemples précis, dans les publications, des problèmes qui se posent lorsqu'il s'agit de déterminer l'année d'étude la plus appropriée pour

l'introduction du matériel; le travail en chimie de Shayer et Adey (1981), cependant, a montré que les auteurs du curriculum et du programme d'études avaient fixé les attentes à des niveaux non réalisables[1]. Si les exigences intellectuelles du curriculum dépassent le potentiel intellectuel de la population étudiante, les conséquences en termes d'évaluation et de rapport sur le rendement des élèves peuvent être graves. C'est ainsi qu'une évaluation peut démontrer la validité, à titre de mesure, des résultats prévus dans le curriculum, mais si les résultats ne sont pas raisonnables, les résultats de l'évaluation ne seront pas une réflexion valide des aptitudes des élèves.

Question concernant la validité : *Les résultats de l'évaluation peuvent-ils être comparés à d'autres résultats d'évaluation en provenance d'autres années d'études ou d'évaluations précédentes?*

Les modèles d'évaluation en lecture et en écriture présentent aussi des variations notables suivant les pays et les époques. Salinger et Campbell (1998) ont fait un rapport sur le développement des évaluations en lecture du *National Assessment of Educational Progress* (NAEP) aux États-Unis. Le test de lecture du NAEP a été administré pour la première fois en 1969 à des échantillons d'élèves représentatifs de tout le pays, le principal mode d'évaluation étant la question à réponse choisie. Cependant le NAEP avait pour mandat de fournir des renseignements sur les tendances du rendement scolaire dans le temps et de faire rapport sur les résultats des élèves par rapport aux objectifs éducationnels courants (Salinger et Campbell, 1998, p. 99). L'évolution des théories et des pratiques sur l'enseignement de la lecture ont amené le NAEP à appliquer deux instruments séparés, l'un qui permet d'étudier les tendances à long terme et est basé sur la façon de procéder de 1969, et un second, fondé sur les théories courantes sur la lecture et l'enseignement de la lecture, ce deuxième instrument faisant de temps en temps l'objet de révisions considérables. C'est pour cette raison que les résultats ne sont pas directement comparables entre les années. Certaines années, notamment en 1992, des rapports séparés ont été publiés pour chaque instrument, une situation qui, de l'avis de Salinger et Campbell, risque de provoquer la confusion dans les médias et parmi les utilisatrices et les utilisateurs des données du NAEP (p. 99).

Ce recours à deux instruments séparés est intéressant en ce qu'il met en évidence la contradiction entre la nécessité de fournir des informations sur les tendances à long terme et la nécessité d'évaluer d'une façon qui soit compatible avec la pratique courante. Bien que les deux instruments fassent un usage étendu des questions à réponse choisie et de l'échantillonnage matriciel, le deuxième instrument utilise un nombre de plus en plus important de questions à réponse élaborée. On remarque d'autres changements dans l'énoncé des objectifs (voir le Tableau 1).

[1] Shayer et Adey, qui travaillaient en Angleterre, ont utilisé une perspective basée sur Piaget pour démontrer que la notion de mole était introduite trop tôt dans le programme de chimie de Nuffield. Leur travail a donné lieu à une révision importante du programme.

TABLEAU 1 – Classification des objectifs du test de lecture du NAEP en 1970-1971 et en 1992

Classification des objectifs de lecture pour le test de 1970-1971 (à titre d'exemple d'instrument à long terme)	Classification des objectifs de lecture pour le test de 1992 (à titre d'exemple d'instrument à long terme)
1. Comprendre ce qui est lu	■ Compréhension initiale
2. Analyser ce qui est lu	■ Élaboration d'une interprétation
3. Utiliser ce qui est lu	■ Réflexion et réaction personnelles
4. Raisonner logiquement à partir de ce qui est lu	■ Démonstration d'esprit critique
5. Porter des jugements sur ce qui est lu	
6. Intérêt pour la lecture et attitudes à son égard	
Source : NAEP (1970)	Source : Council of Chief State School Officers (1992)

C'est la redéfinition de la lecture comme une interaction dynamique et complexe entre trois éléments : la personne qui lit, le texte et le contexte (de l'acte de lecture) (Council of Chief State School Officers, 1992, p. 10) qui est à l'origine de l'inclusion de textes authentiques et de questions non traditionnelles dans les tests. Salinger et Campbell (1998) en décrivent les effets sur le modèle de test :

> L'impact des résultats importants des recherches effectuées pendant les vingt années qui séparent les deux évaluations s'est fait sentir et les experts ont dû reformuler les principes qui sous-tendent l'élaboration, la notation et le rapport des tests. La reformulation tenait compte de la reconnaissance que les lectrices et les lecteurs abordent la lecture avec des connaissances de base différentes et que la lecture n'est en aucune façon une habileté unidimensionnelle simple qui peut être mesurée de façon valide par des questions qui n'ont qu'une seule bonne réponse. En conséquence, le test comprenait des passages plus longs de textes de littérature et d'information reproduits au complet et des « documents » authentiques, comme des horaires d'autobus, qui demandaient une analyse et une interprétation de textes imprimés et non imprimés. Il y avait davantage de questions longues et courtes à réponse élaborée qui requéraient une notation analytique par des équipes d'évaluatrices et d'évaluateurs formés à cet effet. (p.103-104)
>
> [C'est nous qui traduisons.]

L'évaluation des arts du langage en Ontario n'a pas une histoire aussi longue derrière elle et n'a pas encore rencontré les mêmes problèmes que le NAEP, mais on peut appliquer une grande partie de l'expérience de ce dernier à la situation locale, qu'il s'agisse des changements du curriculum ou de la nécessité de recueillir des données de référence.

***Questions concernant la validité** : Quels sont les types d'évaluation les plus appropriés? Quelles sont les caractéristiques de chaque type et du système de notation connexe qui rehaussent ou qui diminuent leur potentiel de généralisation?*

Les changements de la nature des tests de lecture du NAEP correspondent aux orientations adoptées pour les tests administrés

aux élèves de 9e année de l'Ontario en 1992-1993 et pour les évaluations courantes de l'OQRE. La tendance a été de délaisser les tests basés uniquement sur des questions à réponse choisie pour donner la faveur à des réponses élaborées par les élèves et à des tâches authentiques en combinaison avec les questions à réponse choisie. Il est nécessaire, cependant, d'examiner la raison d'être de chaque type de test : les questions à réponse choisie permettent de représenter dans sa totalité le contenu du curriculum alors que les questions à réponse élaborée et les tâches authentiques fournissent des informations directes sur le rendement des élèves.

Les tests d'écriture du NAEP en sont à leur troisième décennie d'utilisation des échantillons d'écriture des élèves pour mesurer le rendement. Les auteurs du cadre utilisé par le NAEP considèrent que ce dernier fait œuvre de pionnier dans les secteurs du recueil d'échantillons réels d'écriture des élèves et de leur notation cohérente (1997, p. ix). Le modèle du test d'écriture de 1998 du NAEP non seulement s'appuie sur son expérience mais incorpore au test les idées issues des cadres exemplaires des États et des recherches récentes sur l'écriture (NAEP, 1997, p. 5). Cette synthèse a donné lieu à la formulation de six objectifs généraux :

- Les élèves doivent écrire à différentes fins : narration, information, persuasion.
- Les élèves doivent effectuer des tâches d'écriture différentes à l'intention de nombreux auditoires différents.
- Les élèves doivent écrire à partir de sources d'inspiration différentes et dans des délais différents.
- Les élèves doivent concevoir, rédiger, réviser et corriger différentes idées et différentes formes d'expression dans leurs écrits.
- Les élèves doivent témoigner du fait qu'elles et ils font des choix efficaces dans l'organisation de leurs écrits. Elles et ils doivent inclure des détails pour illustrer et développer leurs idées et utiliser les conventions appropriées de la langue écrite.
- Les élèves doivent apprécier l'importance de l'écriture comme activité de communication. (1997, p. 5)

Les objectifs guident nécessairement le choix de l'instrument; dans le cas des tests d'écriture, c'est le choix de la raison de l'écrit qui est en cause, p. ex., écrire une lettre pour informer, persuader ou demander. Les éléments retenus par le NAEP sont indiqués dans le Tableau 2. Le Tableau 3 présente une sélection des questions incitatives utilisées par le NAEP pour indiquer l'objectif du discours.

TABLEAU 2 – Éléments à considérer dans la conception des tâches d'écriture (NAEP, 1997, p. 14)

- Buts du discours
  - But principal : narration, information, persuasion
  - Sous-genre : p. ex., exposé de principe, histoire, lettre
- Sujet
  - Source d'information : expérience personnelle, école, nouvelles informations
  - Connaissance du sujet
  - Intérêt
- Complexité intellectuelle
  - Rappeler/Résumer
  - Analyser
  - Déduire/Interpréter
  - Évaluer
- Auditoire
  - Connu/Inconnu
  - Adulte/Enfant
  - Novice/Expert
  - Amical/Hostile
- Présentation
  - Écrite
  - Illustrée
- Critères d'évaluation
- Conditions d'administration
- Processus d'écriture évalués
  - Pré-écriture/Planification
  - Rédaction
  - Révision
  - Correction

TABLEAU 3 – Exemples de tâches d'écriture (NAEP, 1997, p. 47)

	4e année	8e année	12e année
Narration	Fournir des stimuli visuels sur une saison de l'année. Demander aux élèves d'écrire une lettre à l'une ou l'un de leurs grands-parents pour lui raconter une expérience personnelle intéressante en rapport avec la saison.	Fournir des stimuli visuels. Demander aux élèves d'écrire un article à l'intention d'une revue sportive pour raconter l'histoire d'un moment de leur vie où elles ou ils ont pratiqué pour le plaisir un sport ou une activité connexe.	Fournir une citation appropriée. Demander aux élèves d'écrire une lettre à un(e) ami(e) pour lui raconter l'histoire d'un moment de leur vie où elles ou ils ont dû prendre une décision importante.
Information	Fournir une citation appropriée. Demander aux élèves d'expliquer dans une composition à l'intention de leur enseignant(e) de français comment une personne (parent, enseignant(e), ami(e)) les a influencés de façon importante.	Fournir une série de brefs articles de journaux d'une autre période historique. Demander aux élèves d'expliquer ce que cela révèle sur la personne qui a écrit les articles.	Fournir des citations utilisées dans le cadre d'une campagne politique. Demander aux élèves d'en choisir une et de rédiger une composition pour expliquer à leur enseignant(e) de sciences sociales ce qu'elle signifie dans le contexte de la campagne.
Persuasion	Fournir des stimuli visuels sur un animal. Demander aux élèves de convaincre leurs parents ou leurs tuteurs de ce qui constitue le meilleur animal familier.	Fournir, à titre de modèle, de brèves critiques d'un film, d'un programme de télévision ou d'un livre. Demander aux élèves d'écrire une critique pour le journal de l'école qui convaincra d'autres élèves de regarder un film ou un programme de télévision ou de lire un livre qu'elles ou ils aiment particulièrement.	Fournir une citation sur l'éducation aux États-Unis. Demander aux élèves d'écrire une lettre au rédacteur de leur journal local pour prendre position sur un certain aspect de l'éducation et sur leurs propres expériences à cet égard.

L'exemple de tâche d'écriture narrative de 4e année demande aux élèves d'écrire une lettre à l'une ou l'un de leurs grands-parents sur leurs expériences pendant une certaine saison de l'année; il semble donner aux élèves une latitude considérable dans le choix des expériences à raconter. Le sujet, la complexité intellectuelle et l'auditoire sont moins élaborés que pour les élèves de 8e année. Celles-ci et ceux-ci sont invités à rédiger un article à l'intention d'une revue sportive pour raconter l'histoire d'un moment de leur vie où elles et ils se sont adonnés pour le plaisir à un sport ou à une autre activité. On a donc affaire à un sujet basé à la fois sur l'expérience et l'intérêt personnels et sur des aptitudes intellectuelles complexes de rappel et d'évaluation, et l'auditoire est constitué des lectrices et des lecteurs de revues sportives (connues ou inconnues). On assiste à une augmentation similaire de la complexité de chaque type d'écriture pour les années supérieures. Pour avoir un échantillon d'un grand nombre de tâches dans une grande variété de contextes, un minimum de 25 tâches a été proposées pour l'évaluation de 1998 (NAEP, 1997). Chaque élève devait répondre à deux différentes questions incitatives – par exemple de narration et d'information ou de narration et de persuasion – pendant les 50 minutes que durait l'évaluation.

Si, contrairement au mode d'échantillonnage utilisé par le NAEP, l'évaluation a pour objet de fournir des résultats individuels, cela ajoute deux complications notables au modèle à choisir pour le test. La spécificité des tâches (comment généraliser la performance en réponse à cette tâche à d'autres tâches et au curriculum) pose un problème (Linn et Burton, 1994) et, dans le cas des évaluations du Kentucky, il s'est avéré difficile de comparer différents tests (Fairtest, 1998).

Le temps alloué aux évaluations à grande échelle est le résultat d'un subtil équilibre entre le besoin des évaluatrices et des évaluateurs d'obtenir des résultats fiables et le besoin d'enseigner des titulaires de classe qui font pression pour que les conceptrices et les concepteurs produisent des tests qui n'occupent pas une trop grande partie de la journée de classe. Dans le cas d'un test d'écriture basé sur le modèle du NAEP avec trois modes d'écriture, on pourrait demander à chaque élève de ne répondre qu'à une seule question pour chaque mode. S'il n'y a qu'une question pour chaque mode d'écriture, cependant, cela peut produire des distorsions, ce dont on discutera dans la section suivante.

La question de la comparaison entre les tâches est devenue un problème important au Kentucky parce que l'ensemble complet des évaluations fait partie du système de responsabilité de l'État. Comme la structure fait appel à une comparaison ouverte du rendement scolaire d'une année sur l'autre, il est nécessaire de comparer la difficulté des tests. Les problèmes que cela a posés sont décrits de la façon suivante dans *Fairtest Examiner* :

> Du fait qu'il est si facile de se rappeler les quelques questions utilisées chaque année, on ne peut pas les réutiliser sans que certaines et certains élèves ne s'y préparent à l'avance. Mais, si les tâches sont nouvelles chaque année, comment être sûr qu'elles sont d'égale difficulté d'une année sur l'autre? (1998)
>
> [C'est nous qui traduisons.]

La méthode de notation peut aussi avoir un effet sur le potentiel de généralisation des résultats (Hardy, 1992). Les notes des élèves sont rapportées en termes de niveau de rendement basé sur des échelles globales. L'objectif de ces échelles globales est de permettre aux enseignantes et enseignants et aux correctrices et correcteurs d'évaluer une vaste gamme de réponses (Comfort, 1994). Cela n'est pas une tâche aisée. Par exemple, CLASS (1997) offre un exercice de critique des critères des échelles descriptives de notation intitulé *What is wrong here?* « *Il y a quelque chose qui ne vas pas?* » présenté dans le Tableau 4. CLASS montre une rubrique issue d'une étude spéciale du NAEP pour le test d'écriture de 1992 qui utilisait des échantillons des écrits des élèves

rédigés en classe. Les lectrices et les lecteurs étaient invités à examiner la rubrique et à réfléchir aux trois questions critiques du Tableau 4. Les questions de CLASS avaient pour objet de démontrer les incohérences de la rubrique originale du NAEP. Prenons la première question de CLASS – critères énoncés et implicites pour juger la narration; nous voyons que la nature des détails de l'histoire devient un facteur important. Pour une histoire du niveau 2 (non développée), on s'intéresse essentiellement au décor, aux personnages ou aux événements alors qu'aux niveaux supérieurs, par exemple aux niveaux 4 et 5 (histoire construite et développée), on donne de l'importance au décor, aux épisodes, aux buts des personnages ou aux problèmes à résoudre. Dans beaucoup de récits, ce changement subtil des facteurs d'importance peut provoquer un manque de cohérence dans les évaluations.

TABLEAU 4 – Il y a quelque chose qui ne va pas? (CLASS, 1997, p. 15)

Quels sont les critères énoncés et implicites pour juger la narration?

Les élèves pourraient-elles et ils satisfaire à ces critères sans produire d'excellentes narrations?

Quels sont les critères qui ne sont pas mentionnés et devraient l'être?

Guide du NAEP de notation des tests de narration : Écriture – 4e et 8e année

Niveau 1. Description d'événement : La composition est une liste de phrases liées entre elles de façon minimale ou une liste de phrases qui décrivent un seul événement.

Niveau 2. Histoire non développée : La composition est une liste d'épisodes reliés entre eux. Plusieurs événements sont décrits, mais avec peu de détails sur le décor, les personnages ou les événements. (Il n'y a habituellement pas plus d'une phrase sur chaque épisode.)

Niveau 3. Histoire élémentaire : La composition décrit une série d'épisodes et donne des détails (au moins deux ou trois phrases) sur certains aspects de l'histoire (les événements, les buts des personnages ou les problèmes à résoudre). Mais l'histoire manque de cohésion pour des raisons de syntaxe, de développement logique, d'absence de certains événements ou de développement de la conclusion.

Niveau 4. Histoire construite : La composition décrit une suite d'épisodes et donne des détails sur la plupart des éléments de l'histoire (c.-à-d., le décor, les événements, les buts des personnages et les problèmes à résoudre). Mais les histoires sont confuses ou incomplètes (c.-à-d., qu'à la fin, les buts des personnages sont ignorés ou les problèmes ne sont pas réglés comme il faut, que le début ne correspond pas au reste de l'histoire, que la logique interne ou la plausibilité des actions des personnages ne sont pas maintenues).

Niveau 5. Histoire développée : La composition décrit une suite d'épisodes dans laquelle presque tous les éléments de l'histoire sont clairement développés (c.-à-d., le décor, les événements, les buts des personnages et les problèmes à résoudre) et où les objectifs des personnages sont atteints et les problèmes résolus de façon simple à la fin de l'histoire. La composition peut présenter un ou deux problèmes ou donner trop de détails.

Niveau 6. Histoire élaborée : La composition décrit une suite d'épisodes dans laquelle presque tous les éléments de l'histoire sont clairement développés (c.-à-d., le décor, les événements, les buts des personnages et les problèmes à résoudre). La fin présente une résolution élaborée des objectifs ou des problèmes. Les événements présentés et élaborés sont liés entre eux.

Pour 1998, le NAEP a décidé de fournir une description préliminaire des niveaux de rendement sur la base d'une échelle en trois points. Compte tenu du fait que les échantillons d'écriture sont des « premiers jets », la performance des élèves est décrite comme « élémentaire », « compétente » ou « avancée » (NAEP, 1997).

Une autre méthode, la notation dimensionnelle basée sur plusieurs rubriques, vient du désir de spécifité (Pearson, DiStefano et Garcia, 1998). Au lieu d'une seule rubrique, il y en a plusieurs : par exemple, en écriture, il pourrait y avoir le sens de l'auditoire, le style, l'organisation ou la cohérence du contenu, les techniques linguistiques et le point de vue (Pearson et al., 1998). La notation dimensionelle fournit des informations valables, mais Pearson et al. dégagent deux effets négatifs possibles.

Tout d'abord, les enseignantes et les enseignants pourraient insister ouvertement sur certaines dimensions en fournissant à leur sujet des instructions isolées et hors contexte. Ensuite, elles et ils pourraient

> négliger des « sommets » ou des éléments forts de la performance ou ne pas suffisamment en tenir compte. (p. 34)
>
> [C'est nous qui traduisons.]

Pearson et al. insistent sur le fait qu'il est important de reconnaître que les niveaux et les dimensions se chevauchent inévitablement, si bien qu'il y a des chances pour que l'instruction destinée à améliorer la performance dans une dimension l'améliore également dans les autres. (p. 35)

Les rubriques conçues avec soin tiennent compte des besoins de l'évaluatrice ou de l'évaluateur, de l'enseignante ou de l'enseignant et de l'élève, et offrent des occasions importantes de décrire la performance des élèves et de guider les efforts pour améliorer l'apprentissage. Bien que la préparation des rubriques puisse être à la fois exigeante et enrichissante, il incombe à la conceptrice ou au concepteur de réfléchir aux conséquences à la fois positives et négatives de leur utilisation.

Question concernant la validité : *Comment réduire les biais non intentionnels dans l'évaluation?*

La question est complexe et c'est pourquoi nous estimons qu'il vaut mieux parler de réduire que d'éliminer les biais dans les évaluations. Notre point de vue est fondé sur l'étude approfondie de l'équité dans les évaluations de Gipps et Murphy (1994), qui concluent de la façon suivante :

> Les tests justes n'existent pas et ne peuvent pas exister : la situation est trop complexe et la notion simpliste. Cependant, en tenant compte de ce que nous savons sur les facteurs en cause dans les évaluations, leur administration et leur notation, nous pouvons nous efforcer d'élaborer des tests qui soient plus justes envers tous les groupes auxquels ils ont des chances d'être administrés, et ceci est particulièrement important pour les évaluations utilisées à des fins de bilan et de responsabilité. (p. 273-74)
>
> [C'est nous qui traduisons.]

Les facteurs possibles à prendre en compte pourraient inclure le type de question et le langage utilisé en relation avec le sexe, la classe sociale ou le statut socio-économique, les origines ethniques, les possibilités et les expériences d'apprentissage, et d'autres facteurs contextuels.

Eu égard aux types de question, les preuves indiquent que les questions à réponse choisie avantagent les garçons (Walkerdine, 1987) alors que les exercices qui demandent une réponse écrite élaborée avantagent les filles (Black, 1998). Mais cette distinction même n'est pas simple, car Black rapporte que les garçons réussissent mieux dans les compositions sur des sujets impersonnels alors que les filles sont avantagées par les compositions qui ont une portée personnelle et humaine. Beaucoup voient dans les évaluations basées sur la performance une façon d'assurer l'équité entre les sexes, et cet optimisme a récemment été confirmé. Par exemple, dans la Troisième enquête internationale sur les mathématiques et les sciences (TEIMS), les résultats de l'enquête principale, constituée essentiellement de questions à réponse choisie, montraient que les garçons de 8e année dépassaient les filles en sciences de la terre, physique et chimie alors qu'il y avait peu de preuves de différence entre les sexes dans l'évaluation de la performance (Harmon et al., 1997, p. 108).

L'évaluation à partir du dossier d'apprentissage offre des occasions importantes aux élèves de fournir des preuves de leur compétence dans le milieu « sûr » de leur propre classe. Pour les évaluations à grande échelle, beaucoup de choses doivent être prises en considération. Gearhart et Herman (1995) ont écrit le fameux article intitulé *Whose Work Is It?* (« Qui a fait ce travail? ») où, après avoir analysé plusieurs projets d'évaluations à grande échelle, ils dégagent les problèmes associés à l'assignation de notes individuelles pour le travail effectué dans le milieu favorable que représente la salle de classe. Ils ont trouvé que, lorsque les enseignantes et les enseignants avaient activement soutenu l'idée d'une communauté d'apprentissage,

le travail des élèves les plus engagés se retrouvait sous le nom des autres (p. 4). L'expérience du Vermont avec l'utilisation du dossier d'apprentissage pour des évaluations à l'échelle de l'État en mathématiques et en anglais est devenue une légende souvent évoquée dans certains milieux de la communauté d'évaluation. Plus précisément, la fiabilité des notes était très faible et ceci, apparemment pour deux raisons, le manque de cohérence des jugements des enseignantes et des enseignants et le manque d'uniformité dans la composition du dossier (Koretz, Stecher, Klein, McCaffery et Dreibert, 1993). Par contre, LeMahieu, Eresh et Wallace (1995) font rapport sur un projet où un grand nombre des problèmes ont été surmontés. Cependant, le titre de leur article donne à réfléchir. La traduction française en serait : « L'utilisation des dossiers d'apprentissage dans les évaluations à grande échelle : difficile mais pas impossible ». Les publications courantes sur l'usage des portfolios semblent recommander leur utilisation dans la classe (Barton et Collins, 1997; Burger et Burger, 1994; Hebert, 1992) plutôt que pour les évaluations à grande échelle.

Les hypothèses sur lesquelles se fondent les conceptrices et les concepteurs de tests peuvent ne pas correspondre à la vie quotidienne des élèves. Par exemple, Pearson et al. (1998) discutent d'un test de réflexion critique dans lequel une question portait sur une situation concernant les détritus dans un parc :

> Les critères d'évaluation partaient du principe que les détritus étaient inesthétiques mais non dangereux, ce qui témoignait d'une vision suburbaine bourgeoise de ce que représentent les détritus dans un parc. Les critères ne tenaient pas compte de la façon dont on avait appris aux élèves qui traînent dans les rues des quartiers pauvres des villes à réagir devant les milieux dangereux où les détritus abandonnés dans un parc peuvent comprendre des drogues ou des seringues. (p. 42)
>
> [C'est nous qui traduisons.]

Pearson et al. proposent que les conceptrices et les concepteurs de tests visent un type d'équité qui donne aux élèves l'occasion de faire de leur mieux. Ils citent une recherche effectuée précédemment par une personne de leur groupe (Garcia, 1991) qui montrait que les élèves d'anglais langue seconde montrent souvent une meilleure compréhension d'un texte anglais si elles et ils sont autorisés à y répondre – oralement ou par écrit – dans leur langue maternelle (p. 43). Il est certain que cette méthode élargirait l'acceptabilité des tests dans les régions où le nombre d'élèves d'anglais langue seconde est élevé, mais cela ne ferait que déplacer le problème d'équité vers la question du choix des langues utilisées.

Validité des évaluations en mathématiques

Nous examinerons maintenant d'autres problèmes de validité en rapport avec la première évaluation de 3e année effectuée en 1997. Il s'agissait d'une évaluation des habiletés des enfants sur la base du *Programme d'études commun de l'Ontario* en vigueur à l'époque.

Question concernant la validité : *L'instrument choisi correspond-il à la théorie de l'apprentissage qui sous-tend le curriculum à mesurer?*

Au cours de la dernière décennie, les réformatrices et les réformateurs de mathématiques ont suggéré d'apporter des changements radicaux à l'enseignement des mathématiques. Les changements sont basés sur l'opinion que les enfants apprennent en construisant activement leurs propres connaissances mathématiques plutôt qu'en recevant toutes leurs informations de l'enseignante ou de l'enseignant. En 1989, le Conseil national de recherche a exhorté les enseignantes et les enseignants à cesser de voir les mathématiques comme un système rigide de règles dictées de l'extérieur régies par des normes d'exactitude, de vitesse et de mémoire (p. 44) et de modifier leur enseignement en conséquence. Les principes essentiels du mouvement ont été soigneusement définis par le National Teachers Council of Mathematics dans le document intitulé *Standards* (1989) et par l'Ontario Association of Mathematics Educators dans *Focus on Renewal* (1993). Mieux encore, cet effort de réforme a acquis suffisamment de crédibilité pour être adopté par le curriculum de l'Ontario qui place la résolution des problèmes au cœur de l'enseignement des mathématiques.

> Les attitudes, stratégies et méthodes de réflexion que les élèves sont censés acquérir pour résoudre efficacement des problèmes devraient faire partie intégrante de tous les aspects du programme de mathématiques. La résolution de problèmes n'est donc pas traitée comme un domaine distinct [mais] joue plutôt un rôle central dans l'apprentissage des élèves dans chacun des cinq domaines. (Ministère de l'Éducation et de la Formation, 1997, page 65)

Les évaluations complémentaires sont vues par les réformatrices et les réformateurs comme un aspect fondamental à la fois de l'exactitude de la mesure et de l'établissement d'un programme constructiviste, c'est-à-dire basé sur la résolution des problèmes. C'est pourquoi les tenants du mouvement de réforme des mathématiques ont fortement recommandé de remplacer les tests traditionnels avec crayon et papier par des évaluations complémentaires (Kamii et Lewis, 1991; National Council of Teachers of Mathematics, 1995; Ontario Association for Mathematics Educators, 1996; Cobb, Wood et Yackel, 1990; Burns, 1994). Si le programme-cadre de mathématiques considère que la construction des connaissances par l'enfant est plus importante que l'apprentissage par cœur, l'instrument d'évaluation doit être conçu en conséquence. Le modèle d'évaluation doit permettre aux élèves d'acquérir « un pouvoir mathématique » en développant leur habileté à analyser, à faire des hypothèses, et à raisonner de façon logique, ainsi que leur aptitude à utiliser une variété de méthodes mathématiques pour résoudre des problèmes différents (National Council of Teachers of Mathematics, 1995, p. 5).

Validité indirecte de l'évaluation comme instrument de réforme

Y a-t-il une preuve que l'évaluation a donné lieu ou donnera lieu à une amélioration des méthodes d'enseignement et donc à une amélioration du rendement scolaire?

Dans la plupart des évaluations à grande échelle basées sur la performance, on part du principe que, si les enseignantes et les enseignants reçoivent une évaluation de l'extérieur des résultats de leurs élèves et qu'elles et ils sont exposés à des tâches d'évaluation exemplaires, leur enseignement s'améliorera et les résultats des élèves s'amélioreront. Mais Hansen (1993) a analysé les publications américaines sur les rapports entre l'obligation de rendre compte et la réforme de l'éducation au cours des dernières décennies et a conclu que les preuves de l'efficacité de l'obligation de rendre compte comme instrument de réforme étaient extrêmement faibles (p. 11). Il faut reconnaître qu'une grande partie de cette recherche est basée plutôt sur des tests normalisés plus traditionnels que sur des évaluations basées sur la performance. L'auteur conclut que l'évaluation basée sur la performance peut être efficace si elle fait partie d'une réforme systémique qui inclut un perfectionnement du personnel et une aide technique axés sur l'unité visée par le changement – l'école individuelle (p. 19). Baron (1996), analysant l'expérience du Connecticut, convient qu'il y a une place pour l'évaluation à grande échelle basée sur la performance.

> Les preuves indiquent de plus en plus que, lorsqu'on la compare aux formes d'évaluation plus traditionnelles, l'évaluation basée sur la performance a davantage de chances d'augmenter l'accès et de renforcer la capacité des enseignantes et des enseignants et des élèves en offrant des modèles d'apprentissage et des activités d'évaluation qui correspondent aux pratiques définies dans une variété de disciplines. (p. 189)
>
> [C'est nous qui traduisons.]

D'autres, comme Worthen (1993), font écho à cet optimisme prudent :

> J'espère que les spécialistes de l'évaluation et les autres éducatrices et éducateurs réfléchiront à ces problèmes et aux questions connexes de façon à pouvoir dégager, et éventuellement réaliser, toutes les possibilités de l'évaluation complémentaire. (p. 453)
>
> [C'est nous qui traduisons.]

L'étude de cas effectuée par Firestone, Mayrowetz et Fairman (1998) sur les effets des évaluations de mathématiques à grande échelle basées sur la performance dans cinq districts scolaires du Maryland et du Maine a révélé des changements importants dans l'alignement du contenu d'enseignement sur les questions des tests mais n'ont guère trouvé de preuve de changement dans les méthodes pédagogiques. Certaines chercheuses et certains chercheurs suggèrent que cela risque de devenir la norme, au moins en ce qui concerne l'enseignement des mathématiques (Cohen et Ball, 1990; Cohen, 1995). Elles et ils estiment que la faiblesse de l'enseignement des mathématiques est due à la conviction profonde des enseignantes et des enseignants que les mathématiques ne sont rien d'autre que l'application de règles apprises par cœur ainsi qu'aux limitations de leurs propres connaissances (Battista, 1994; Frank, 1988; Firestone et al., 1998; Schifter et Fosnot, 1993). Il s'ensuit que, bien que les évaluations authentiques puissent indiquer aux enseignantes et aux enseignants les insuffisances de leurs méthodes d'enseignement, il en faut beaucoup plus pour les amener à passer à l'étape suivante de la réforme des mathématiques. De ce point de vue, les enseignantes et les enseignants ont besoin de consacrer du temps à réapprendre les mathématiques d'une façon constructiviste et basée sur les problèmes afin d'être en mesure d'enseigner de cette manière. La réforme authentique de la pratique des enseignantes et des enseignants de mathématiques, une réforme qui pourrait permettre aux élèves de mieux comprendre les mathématiques, pose un défi de taille.

Validité indirecte de l'évaluation comme instrument de réforme

L'importance de l'évaluation basée sur la performance dans la réforme de l'enseignement est due aux renseignements nouveaux qu'elle donne aux enseignantes et enseignants sur le véritable niveau de compréhension de leurs élèves. Ces évaluations risquent d'amener les enseignantes et les enseignants à penser les mathématiques d'une façon neuve basée sur leur expérience d'un enseignement à base constructiviste.

Les avantages risquent d'être moindres, cependant, si les tests sont préparés en secret, communiqués aux écoles, puis envoyés à l'extérieur pour les faire noter par des machines (Darling-Hammond et Ancess, 1996). C'est généralement ce qui s'est passé dans le cas des tests basés sur des questions à réponse choisie, mais Darling-Hammond et Ancess craignent que cette façon de procéder ne limite les possibilités de réformes à partir des évaluations basées sur la performance.

> Les évaluations qui sont élaborées et notées à l'extérieur n'ont guère de chances de transformer les connaissances et la compréhension des enseignantes et des enseignants – et des organisations scolaires – même si elles sont davantage basées sur la performance que les tests courants. Cela est lié au fait que la réflexion des enseignantes et des enseignants sur les structures plus profondes du curriculum, la nature et les nuances de la pensée des élèves et les rapports entre les efforts d'apprentissage et la performance des élèves dérive essentiellement d'un contact constructiviste de première main avec l'élaboration du test et l'évaluation subséquente du travail des élèves. (p. 53)
>
> [C'est nous qui traduisons.]

Darling-Hammond et Ancess font valoir que l'évaluation devrait non seulement produire un résultat numérique quelques mois plus tard, mais constituer un sujet de conversation continu avec les parents et les autres enseignantes et enseignants (p. 76). Il faut au moins que les tâches soient disponibles à titre d'exemples pour aider les enseignantes et les enseignants à élaborer leurs propres évaluations.

La valeur à long terme des évaluations à grande échelle comme instruments de réforme reste à déterminer. Baron (1996) indique que les résultats du Connecticut témoignent en leur faveur et nous demande de ne pas oublier que, si on leur accorde l'attention et le soutien nécessaires, ces évaluations peuvent fortement contribuer à améliorer les résultats des enseignantes et des enseignants et des élèves. Il ne faut cependant pas y voir une panacée (p. 189).

Revenons, en conclusion, sur les nouvelles évaluations pratiquées en Ontario. Les tests récents de l'OQRE poussent plus loin les méthodes, recommandées par la Commission royale sur l'éducation, qui offrent des expériences d'évaluation authentiques aux élèves et à leurs enseignantes et enseignants tout en satisfaisant aux exigences de responsabilité en fournissant les données désirées aux parents et aux contribuables. Nous estimons que ces évaluations représentent une orientation positive pour l'évaluation du rendement scolaire en Ontario, mais il faut fournir des preuves claires de la « valeur fonctionnelle » (Messick, 1989) de ce programme d'évaluation. Nous proposons qu'une étude approfondie de la validation soit effectuée pour examiner cette question.

Une étude approfondie de la validation entraînerait un examen sérieux d'un grand nombre des caractéristiques présentées dans le présent article, par exemple, le rôle des théories de l'apprentissage sur le choix du modèle d'évaluation. Pour les évaluations basées sur la performance, les questions techniques de validité et de fiabilité requièrent une analyse approfondie allant au-delà des approches statistiques qui sont depuis plusieurs décennies à l'origine des tests traditionnels. Ces approches ont une certaine valeur mais elles ne seraient pas les seules sources d'information dans l'étude de la validation. En particulier, il y aurait lieu de faire des recherches étendues et consciencieuses au niveau de la salle de classe pour déterminer dans quelles circonstances on a assisté à des développements positifs et ce qui pourrait être fait pour étendre les effets positifs et réduire les conséquences négatives.

Bibliographie

AERA, APA et NCME (American Educational Research Association, American Psychological Association et National Council on Measurement in Education). *Standards for educational and psychological testing*. Washington, DC, American Psychological Association, 1985.

Baron, J. B. « Developing performance-based student assessments: The Connecticut experience ». Dans J. Baron et D. Wolf (éd.), *Performance-based student assessment: Challenges and possibilities*. Chicago, University of Chicago Press, 1996, p. 166-191.

Barton, J. et A. Collins. *Portfolio assessment: A handbook for educators*. Menlo Park, Calif., Addison Wesley, 1997.

Battista, M. « Teachers' beliefs and the reform movement in mathematics education ». *Phi Delta Kappan*, 75 (1994), 464-470.

Black, P. *Testing: Friend or foe?* London, Falmer Press, 1998.

Brown, M., McCallum, B., Taggart, B. et C. Gipps. « The validity of national testing at age 11: The teacher's view ». *Assessment in Education*, 4(2) (1997), 271-293.

Burger, S. et D. L. Burger. « Determining the validity of performance-based assessments ». *Educational Measurement: Issues and Practice, 13*(1) (1994), 9-15.

Burns, M. « Arithmetic: The last holdout ». *Phi Delta Kappan*, (février 1994), p. 471-476.

Cannell, J. « National norm-referenced elementary achievement testing in America's public schools: How all fifty states are above the national average ». *Educational Measurement: Issues and Practice*, 7(2) (1988), 5-9.

CLASS. *Rubrics and scoring criteria*. Pennington, N.J., CLASS, 1997.

Cobb, P., Wood, T. et E. Yackel. « Classrooms as learning environments for teachers and researchers ». *Journal for Research in Mathematics Education Monograph No. 4*, (1990), 125-146.

Cohen, D. K. « What is the system in systemic reform? ». *Educational Researcher*, 24(9) (1995), 11-17.

Cohen, D. et D. Ball. « Relations between policy and practice: A commentary ». *Educational Evaluation and Policy*, 12(3), (1990), 249-256.

Cole, N. S. « The impact of science assessment on classroom practice ». Dans G. Kulm et S. Malcolm (éd.), *Science assessment in the service of reform*. Washington, D.C., American Association for the Advancement of Science, 1991, p. 97-105.

Comfort, K. B. « A *sampler of science assessment: Elementary* ». Sacramento, California Department of Education, 1994.

Commission royale sur l'éducation de l'Ontario. *Pour l'amour d'apprendre*. Toronto, Imprimeur de la Reine, 1994.

Council of Chief State School Officers, 1992. Cité dans Salinger et Campbell, 1998.

Darling-Hammond, L. et J. Ancess. « Authentic assessment and school development ». Dans J. Baron et D. Wolf (éd.), *Performance-based student assessment: Challenges and possibilities*. Chicago, University of Chicago Press, 1996, p. 52-83.

Fairtest. « Equating was real problem with Kentucky performance events ». *Fairtest Examiner*, (été 1998). (http://www.fairtest.org/examarts/summer98/k-ky-2.htm).

Feldt, L. S. et R. L. Brennan. « Reliability ». Dans R. L. Linn (éd.), *Educational measurement*, 3e éd. New York, Macmillan, 1989, p. 105-146.

Bibliographie

Firestone, W., Mayrowetz, D. et J. Fairman. « Performance-based assessment and instructional change: The effects of testing in Maine and Maryland ». *Educational Evaluation and Policy Analysis*, 20(2) (1998), 95-113.

Frank, J. « Problem solving and mathematical beliefs ». *Arithmetic Teacher*, (janvier 1988), p. 32-34.

Garcia, G. E. (1991). Cité dans Pearson et al., 1998.

Gearhart, M. et J. L. Herman. « Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability ». *Evaluation Comment*, (hiver 1995), p. 1-16.

Gipps, C. *Beyond testing: Towards a theory of educational assessment*. London, Falmer Press, 1994.

Gipps, C. et P. Murphy. *A fair test*. Buckingham, UK, Open University Press, 1994.

Hansen, J. « Is educational reform through mandated accountability an oxymoron? ». *Measurement and Evaluation in Counseling and Development*, 26 (avril 1993), 11-21.

Hardy, R. *Options for scoring performance assessment tasks*. Communication présentée à l'assemblée annuelle du National Council on Measurement in Education, San Francisco, avril 1992.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez, E. J. et G. Orpwood. *Performance assessment in IEA's Third International Mathematics and Science Study*. Boston, Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1997.

Hebert, E. A. « Portfolios invite reflection from students and staff ». *Educational Leadership*, 49(8) (1992), 58-61.

Herman, J., Aschbacher, P. et L. Winters. *A practical guide to alternative assessment*. Alexandria, VA, ASCD, 1992.

Horner, S. « Assessing reading in the English National Curriculum ». Dans C. Harrison et T. Salinger (éd.), *Assessing reading 1: Theory and practice*. London, Routledge, 1998, p. 84-95.

Jones, K. et B. L. Whitford. « Kentucky's conflicting reform principles ». *Phi Delta Kappan*, 79(4) (décembre 1997), 276-281.

Kamii, C. et B. A. Lewis. « Achievement tests in primary mathematics: Perpetuation of lower order thinking ». *Arithmetic Teacher*, (mai 1991), p. 4-9.

Kirst, M. et C. Mazzeo. « The rise, fall, and rise of state assessment in California ». *Phi Delta Kappan*, 78(4) (décembre 1996), 319-323.

Koretz, D., Klein, S., McCaffery, D. et B. Stecher. *Interim report: The reliability of the Vermont portfolio scores in the 1992-93 school year*. Santa Barbara, CA., RAND Institute on Education and Training, 1993.

Koretz, D., Stecher, B., Klein, S., McCaffery, D. et E. Deibert. *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont experience*. RAND Institute on Education and Training, 1992.

LeMahieu, P., Eresh, J. et R. Wallace. « Portfolios in large-scale assessments: Difficult but not impossible ». *Educational Measurement: Issues and Practice*, 14(5) (automne 1995), 11-16, 25-28.

Linn, R. L. et E. Burton. « Performance-based assessment: Implications of task specificity ». *Educational Measurement: Issues and Practice*, 13(1) (1994), 5-8, 15.

Lovitts, B. E. et A. B. Champagne. « Assessment and instruction: two sides of the same coin ». Dans A. B. Champagne, B. E. Lovitts et B. J. Callinger (éd.), *This year in school science 1990: Assessment in the service of instruction*. Washington, DC, American Association for the Advancement of Science, 1990, p. 1-13.

Bibliographie

Madaus, G. et T. Kellaghan. « The British experience with "authentic" testing ». *Phi Delta Kappan*, (février 1993), p. 358-469.

Mehrens, W. « Using performance assessment for accountability purposes ». *Educational Measurement: Issues and Practice*, 11(1), (1992), p. 3-9.

Messick, S. « Validity ». Dans R. L. Linn (éd.), Educational Measurement, 3e éd. New York, Macmillan 1989, p. 13-103.

Ministère de l'Éducation et de la Formation de l'Ontario. *Le curriculum de l'Ontario, de la 1re à la 8e année, Français*. Toronto, Imprimeur de la Reine, 1997.

Ministère de l'Éducation et de la Formation de l'Ontario. *Le curriculum de l'Ontario, de la 1re à la 8e année, Mathématiques*. Toronto, Imprimeur de la Reine, 1997.

Ministère de l'Éducation et de la Formation de l'Ontario. *Le programme d'études commun : Politiques et résultats d'apprentissage, de la 1re à la 9e année*. Toronto, Imprimeur de la Reine, 1995.

NAEP. *Voir* National Assessment of Educational Progress.

National Assessment of Educational Progress, 1970. Cité dans Salinger et Campbell (1998).

National Assessment of Educational Progress. *Writing framework and specifications for the 1998 National Assessment of Educational Progress*. Washington, DC, National Assessment Governing Board, 1997.

National Council of Teachers of Mathematics. *Assessment standards for school mathematics*. Reston, VA, NCTM, 1995.

National Council of Teachers of Mathematics. *Curriculum and evaluation standards for school mathematics*. Reston, VA, NCTM, 1989.

National Research Council. *Everybody counts: A report to the nation on the future of mathematics teaching*. Washington, DC, National Academy Press, 1989.

Nolen, S., Haladyna, T. et N. Hass. « Uses and abuses of achievement test scores ». *Educational Measurement: Issues and Practice*, (été 1992), p. 9-15.

Office de la qualité et de la responsabilité en éducation. *Rapport provincial sur le rendement*. Toronto, Imprimeur de la Reine, 1997.

Ontario Association for Mathematics Educators. *Linking assessment and instruction in mathematics*. Toronto, OAME, 1996.

Ontario Association for Mathematics Educators. *Focus on renewal of mathematics*. Toronto, OAME, 1993.

Pearson, P. D., DiStefano, L. et G. E. Garcia. « Ten dilemmas of performance assessment ». Dans C. Harrison et T. Salinger (éd.), *Assessing reading 1: Theory and practice*. London, Routledge, 1998, p. 21-49.

Raymond, M. R. et W. M. Houston, *Detecting and correcting for rater effects in performance assessment*. Communication présentée aux assemblées annuelles de l'American Educational Research Association et du National Council on Measurement in Education, Boston, avril 1990.

Salinger, T. et J. Campbell. « The national assessment of reading in the USA ». Dans C. Harrison et T. Salinger (éd.), *Assessing reading 1: Theory and practice*. London, Routledge, 1998, p. 96-109.

Schifter, D. et C. Fosnot. *Reconstructing mathematics education: Stories of teachers meeting the challenge of reform*. New York, Teachers College Press, 1993.

Shavelson, R. J., Gao, X. et G. P. Baxter. « On the content validity of performance assessments: Centrality of domain specification ». Communication sollicitée pour la *First European Electronic Conference on Assessment and Evaluation*. Conférence par courrier électronique de l'European Association for Research on Learning and Instruction (EARLI), 1994.

Bibliographie

Shayer, M. et P. Adey. *Toward a science of science teaching: cognitive development and curriculum demand.* London, Heinemann, 1981.

Shepard, L. « Why we need better assessments ». *Educational Leadership, 45* (avril 1989), 4-9.

Slater, T. F. et J. M. Ryan. « Laboratory performance assessment ». *The Physics Teacher, 31* (1993), 306-308.

Smith, M. L. « Put to the test: The effects of external testing on teachers ». *Educational Research,* (juin/juillet 1991), p. 8-11.

Stake, R. « Some comments on assessment in U.S. education ». *Educational Policy Analysis Archives, 6*(14) (juillet 1998). (http://olam.ed.asu.edu/epaa/v6n14.html)

Traub, R. E. et G. L. Rowley. « Understanding reliability ». *Educational Measurement: Issues and Practice, 10*(1) (1991), 37-45.

Tyler, R. *Basic principles of curriculum development.* Chicago, University of Chicago Press, 1949.

Walkerdine, V. « Some issues in the historical construction of the scientific truth about girls ». Dans A. Kelly (éd.), *Science for girls.* Milton Keynes, UK, Open University Press, 1987, p. 37-44.

Wiggins, G. *Assessing student performance.* San Francisco, Jossey-Bass, 1993.

Wiggins, G. « A true test: Toward more authentic and equitable assessment ». *Phi Delta Kappan, 71*(9) (mai 1989), 703-713.

Worthen, B. « Critical issues that will determine the future of alternative assessment ». *Phi Delta Kappan,* (février 1993), p. 444-454.

Les auteur(e)s

Anthony Bartley est professeur adjoint à la faculté des sciences de l'éducation de l'Université Lakehead à Thunder Bay. Son enseignement et ses recherches sont axés sur l'enseignement des sciences, les nouvelles formules d'évaluation et l'application de la théorie de la validité aux évaluations à grande échelle.

Alexandra Lawson est chargée de cours sur l'enseignement des mathématiques au palier élémentaire dans le programme de formation permanente du personnel enseignant de l'Institut d'études pédagogiques de l'Ontario de l'Université de Toronto. Elle prépare un doctorat dans ce domaine et s'intéresse particulièrement à l'impact de l'évaluation sur l'enseignement des mathématiques.